Who Teaches the Teachers? a Rct of Peer-to-Peer Observation and Feedback in 181 Schools

Gillian Wyness (University College London)
Richard Murphy (University of Texas at Austin)
Felix Weinhardt (DIW Berlin)

Discussion Paper No. 116

September 13, 2018

Collaborative Research Center Transregio 190 | www.rationality-and-competition.de
Ludwig-Maximilians-Universität München | Humboldt-Universität zu Berlin
Spokesperson: Prof. Dr. Klaus M. Schmidt, University of Munich, 80539 Munich, Germany
+49 (89) 2180 3405 | info@rationality-and-competition.de

Who teaches the teachers? A RCT of peer-to-peer observation and feedback in 181 schools*

Richard Murphy[†] Felix Weinhardt[‡] Gill Wyness[§]

8th August 2018

Abstract

It is well established that teachers are the most important in-school factor in determining student outcomes. However, to date there is scant robust quantitative research demonstrating that teacher training programs can have lasting impacts on student test scores. To address this gap, we conduct and evaluate a teacher peer-to-peer observation and feedback program under Randomized Control Trial (RCT) conditions. Half of 181 volunteer primary schools in England were randomly selected to participate in the two year program. We find that students of treated teachers perform no better on national tests a year after the program ended. The absence of external observers and incentives in our program may explain the contrast of these results with the small body of work which shows a positive influence of teacher observation and feedback on pupil outcomes.

JEL codes: I21, I28, M53

Keywords: education, teachers, RCT, peer mentoring

*We thank Stephen Machin, Chris Karbownik, Anna Raute, and Eric Taylor for valuable feedback and comments, as well as participants of the Bonn/BRIC Economics of Education Conference, the Manheim labour seminar and of the IWAEE. We thank the UK Department for Education for access to the English student census data under DR160317.03. Weinhardt gratefully acknowledges financial support by the German Research Foundation through CRC TRR 190. All errors are our own.

[†]University of Texas at Austin, NBER, IZA, CESifo and CEP at the LSE.

[‡]DIW Berlin, IZA, CESifo and CEP at the LSE.

[§]UCL Institute of Education and CEP at LSE.

1 Introduction

It is well established that teachers are the most important in-school factor in determining student outcomes (Rockoff, 2004; Rivkin et al., 2005). Thus, the fact that there is huge variation in teacher quality (Hanushek and Rivkin, 2010), is a perennial problem for education policy-makers. One obvious course of action would be to hire better teachers; however, many studies have concluded that teacher effectiveness is very difficult to predict from teacher characteristics (Aaronson et al., 2007; Kane et al., 2008) reducing the viability of this solution. An alternative would be to simply dismiss poorly performing teachers (Hanushek and Rivkin, 2010; Chetty et al., 2014), but this too is a challenge given the administrative burden required, difficulties with replacements and lack of good information on teacher effectiveness available to school principals (Jacob et al., 2016; Rothstein, 2015).

Consequently, a potentially powerful strategy for policy-makers concerned with improving educational outcomes would be to improve the quality of the stock of existing teachers either through incentives or teacher training programs. Research in this area has tended to focus on the former, with a number of studies evaluating the use of performance related pay as a means to improve teacher productivity (Lavy, 2009; Goodman and Turner, 2010; Springer et al., 2011; Muralidharan and Sundararaman, 2011; Neal, 2011). However, these studies have had mixed results, calling into question the effectiveness of performance related pay as a magic bullet to improve educational outcomes in developed countries. An alternative means of improving teacher performance on-the-job, and the subject of this paper, is through teacher training programs. Most recently, Taylor and Tyler (2012) find positive evidence on the effectiveness of one particular type of teacher development -teacher feedback. However, in summary there exists little robust quantitative research demonstrating that teacher training programs can have lasting impacts on student test scores.[1]

This study estimates the causal effect of teacher peer-to-peer mentoring on student outcomes under RCT conditions, randomised at the school level. In the program studied here, fourth and fifth grade teachers work in small groups of three to plan lessons that address shared teaching and learning goals. They then observe

[1]We provide a detailed review of the literature at the end of this section.

each other's lessons, provide feedback, and refine future lesson plans. Each year each teacher is observed three times by her two peers. This process is repeated throughout a two year period for a total of eighteen lesson observations. To ensure structured feedback and implemention, all participating teachers received five full training days held by educational experts on teaching mentoring.

There are a number of reasons why we may expect peer-to-peer mentoring to be an effective form of teacher training. Unlike many other professions, teachers do not interact with their peers in the classroom. Thus, classroom observations offer an opportunity for teachers to see, and be seen in action. Feedback on their observed performance could thus provide teachers with new detailed information on their performance in the classroom. Given that teachers have been shown to be motivated agents (Dixit, 2002), this could result in improved planning and preparation and subsequently better performance (Steinberg and Sartain, 2015). It could encourage teachers to self-reflect and attempt to acquire work-related skills as a result of their peers influence (Jackson and Bruegmann, 2009), and could lead to discussion with other teachers resulting in improvements in teaching practice across the school as a whole (Taylor and Tyler, 2012).

Perhaps unsurprisingly then, many schools carry out peer observation programs informally, albeit with little instruction or consistency (Weisberg et al., 2009), making them difficult to evaluate empirically. Moreover, testing the impact of teacher observation, and teacher training in general, on pupil outcomes is an empirical challenge due to non-random selection of teachers (and students) into training. Our trial is large-scale, with 543 teachers teaching a total of 13,000 students, over two cohorts in all subjects, across 181 primary schools in England. Despite having strict experimental conditions, our experiment is conducted within schools, in a manner which could easily be replicated or taken to scale. Thus, we capture the impact of teacher observation and feedback in a 'real-world' setting.

Our outcomes of interest come from national, compulsory, high stakes, externally marked academic tests intended to measure student learning throughout primary school. The tests are conducted at the end of primary school in sixth grade, when the pupil is aged 10/11, one year after the intervention. As such, our study does not suffer from any biases associated with tests implemented by the school or teachers or trainers themselves. A further benefit of using national tests is that we can exploit administrative data linked to the pupil's outcomes at

both the treatment schools and the control schools. As a result, we did not need to contact the control schools again or do any testing in these schools, after they were informed that they were not selected to receive the treatment. Thus, our setting allows us to provide new and highly compelling evidence on the impact of teacher observation and feedback on pupil outcomes.

In summary, we find no evidence that teacher peer observation and feedback increases pupil performance compared to business as usual in the classroom. Because of our unique setting, we can control for previous test scores at the student level, and exploit the panel nature of the administrative data to estimate a RCT-difference-in-differences analysis. We can reject positive effects on student test scores of about ten percent of a standard deviation across all subjects, and effect sizes larger than five percent of a standard deviation in reading and writing tests.

This study is directly related to the literature on teacher training and student outcomes, which mainly uses quasi-experimental methods to estimate causal effects, for example Jacob and Lefgren (2008), Harris and Sass (2011), and Angrist and Lavy (2001) with only the latter finding a positive effect. Experimental evaluations of training programs have also failed to find any impact (Garet et al., 2010, 2011). However, none of these studies examine teacher peer-to-peer observation as a form of teacher development.

One quasi-experimental study by Taylor and Tyler (2012), directly examines effects of teacher observation on student performance and does find a significant positive impact. In this setting, teachers in Cincinnati Public Schools participated in a year-long classroom observation program known as the Teacher Evaluation System (TES).[2] This program involved three unannounced observations by external experts, and one by the school principal, and involved the provision of formal written feedback and grades to the observed teachers. Identification is based on near-random timing of the implentation of the year long program at a school. The study finds that the students of teachers who have been evaluated improve their maths scores by 11 percent of a standard deviation in the year after the subjective evaluation, and about 16 percent of a standard deviation two years later, compared

[2]Papay et al. (2018) currently have an ongoing teacher observation RCT in the field with an end date of 2020. A pilot study by Steinberg and Sartain (2015) evaluates the Chicago Excellence in Teaching Project (EIP) in which teachers are observed by their principal during a lesson, followed by a feedback session as well as more formal ratings, finds no significant effect.

to students of non-evaluated teachers.

Apart from the RCT conditions, there are a number of important additional differences which differentiate our study from the Cincinnati study and may explain the differing results, and offer mechanisms through which the Cincinnati study generates a positive finding. First, and most notably, our program does not involve teacher incentives; a key tenet of the program is to facilitate free and open discussion, intended to improve the teachers' future performance. Therefore no formal scoring or further consequences are associated with the observations. Conversely, Cincinnati teachers are formally scored, with the results carrying explicit consequences, including impact on promotions and tenure, and potential non-renewal of the teacher's contract. Second, for reasons of scalability, our peer-to-peer observation and feedback program relies solely on existing teachers within the school, while external teachers were responsible for observing and evaluating teachers in the Cincinnati trial. It may well be the case that peers provide less useful feedback compared to external experts. Moreover, the presence of experts in the Cincinnati trial may have created a more formal atmosphere in the classroom, particularly since lessons were also filmed. Finally, while the Cincinnati study is reliant upon quasi-random timing of schools implementing the TES for identification, our study was designed as an RCT, with random assignment of schools to the program along with with a pre-registered statistical analysis plan.

The findings of our study are important. This paper provides the first experimental evaluation of a teacher observation program that is purely designed for teacher development, rather than also including incentives. Our results show that teacher observation and feedback cannot solve the policy maker's problem of huge variation in teacher effectiveness.

The remainder of the paper proceeds as follows: Section 2 provides further details about the intervention. In Section 3 we describe the data used in the analysis, with the RCT design described in section 4. Results are presented in section 5, with a discussion of potential reasons for differences with the existing literature in section 6. Conclusions follow in section 7.

2 The peer-to-peer observation and feedback intervention

2.1 Details of intervention

Teacher peer-to-peer observation and feedback is a type of professional development program with a long history of use in Japan and is increasingly used in the US and worldwide.[3] In this particular program teachers work in small groups to plan lessons that address shared teaching and learning goals.

Specifcally teachers within a schoool form a group of three (known as a learning tripod), with one of the three selected as the expert teacher'. Schools are free to choose which teachers are involved in the intervention, and who would be the expert teacher (though all schools chose teachers with some subject expertise in English or maths as the expert), with the restriction that two of the teachers should be teaching year groups 4 and 5.[4] Training consisted of five full training days for teachers participating in the program. This was conducted by experts in the program and included information on the ethos, protocols and practice. Four of the five training days occurred during the first year. The fifth training day, at the beginning of the second year, was focused on optimising feedback and sustaining the program through its second year. Thus, while the program lasted for two years -and potentially changed teacher practice and student learning for much longer-, the treatment intervention was heavily concentrated in the first year.

The trial was pre-registered with the American Economic Association's registry for RCTs and a detailed statistical analysis plan was approved before we had access to the administrative student outcomes data.[5] The program was delivered

[3]For example Lesson Study Alliance helps US teachers , mainly based in Chicago, use Lesson Study, a peer-to-peer observation and feedback program. See http://www.lsalliance.org/; (Fernandez et al., 2003) study a USJapan lesson study collaboration; Perry and Lewis (2009) describe the use of Lesson Study in a medium-sized California K-8 school district.

[4]Some of the smallest schools had mixed-age classes, and so one teacher may have taught both Year 4 and Year 5. Given the tripod design, if a school had only one class per year group or less they would have to choose a teacher from another year, which was seen as unproblematic from the developer's perspective since the approach does not propose to develop teaching skills specific to a particular year group. We placed no restriction on what other year group was chosen. Because the randomization and analysis is at the school level we are not concerned about schools being able to choose the teachers.

[5]The AEA trial registration number is 1779, for details see: `http://www.socialscienceregistry.org/trials/1779`. The statistical pre-analysis plan can be ac-

independently of this impact evaluation by a team at Edge Hill University with support from external consultants.[6]

The implementation of the program in schools starts with an initial group meeting where the three teachers plan the order in which they are to be observed and which lessons will be observed. The first teacher then teachers her three research lessons' observed by the other two teachers. During these classes, the observing teachers do not interact with students or the teacher but remained solely in their observing role. After each class the group meets to discuss the lesson and plan the next in terms of content, structure and delivery. Over the course of the academic year there were three cycles of the program with each teacher taking the turn of being observed.

The lack of formal scoring highlights that the program's intention is to provide a space for non-judgemental discussion in the school day, rather than a formal program incorporating consequences or incentives.

As mentioned previously, control schools did not receive the treatment at any stage, nor did they receive any information about the treatment or training materials. Thus it is assumed that business as usual conditions applied in these schools.

As discussed, it is plausible that this structured cycle of teacher peer-to-peer observation and feedback would have a positive impact on pupils' educational outcomes. Through the program's cycle, teachers learn new information about their performance from the feedback of the observers, the subsequent conversations taking place between the three teachers, as well as through their own self-reflection. This new information should help them to develop new skills and improve their effectiveness, in particular because teacher training in the UK has very few on-the-job elements. Similar to teachers elsewhere, English teachers receive very little on-the-job feedback or structured opportunities for on-the-job learning once they completed their original qualification.

Since teacher improvement through observation could affect pupil performance in many areas, we estimate the impact of the program on all tested subjects at the end of primary school. These are maths, reading, Spelling Punctuation and Grammar (SPAG) and science. Our pre-specified main outcome of interest is the

cessed here: `https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Protocols/Round_4-Lesson_study_SAP.pdf`

[6]See https://everychildcounts.edgehill.ac.uk/special-projects/lesson-study/ for more details.

students mean performance in reading and maths.

2.2 Timing of intervention

The teacher observation program took place in state primary schools in England[7] during the 2013/14-2015/16 academic years. Figure 1 shows the affected cohorts given the timing of the intervention in calendar years and the target in terms of academic years. In this paper, we analyse effects on age-11 outcomes for two cohorts, which were affected by one (cohort 1) or two years (cohort 2) of this intervention, both measured one year after the end of the intervention, and almost two (cohort 1) or three years (cohort 2) after its start.

[Figure 1 goes here]

3 Setting and data

3.1 Administrative student census data

Our analysis relies on linked-in administrative data that are available for all students in state-education in England from the National Pupil Database (NPD) throughout this time period. In England, pupils attend primary school from age 4/5 to 10/11, taking them from Reception through to Year 6. Pupils take national compulsory tests in Year 2 at age 6/7 (known as Key Stage 1 (KS1)) and at the end of primary school in Year 6 at age 10/11, known as Key Stage 2 (KS2). From now on we refer to these tests as age-7 and age-11 tests, see Figure 1.

The administrative age-7 tests serve as the baseline measure of student achievement. Each student is assessed in math and reading by their teacher and are assigned an achievement level, which takes values between three and 27. Since these national tests are available for all students, we use the mean reading and maths achievement level as measure for initial student ability.

The age 11 tests examine the students ability in four different areas, maths, reading, Spelling Punctuation and Grammar (SPAG) and science. The first three

[7] 93 percent of pupils attend state primary schools in England (DfE, 2015)

of these are externally marked on a 100 point scale, which we percentalise at the national subject-cohort level to ensure comparability across subjects and years. This is important given the national age-11 assessment changed between the first and second cohorts. The exception to this is Science, which is assesed by the teacher and is only reported in 13 coarse levels which makes it inappropriate to be percentalised. Moreover, there is no science outcome for the second cohort as it was not recorded in 2015/16.

Our use of the administrative test score data has four key advantages. First, this data is available for all students and schools with no attrition from the data in the treatment or control groups. Second, we have a comparable measure of student achievment prior to the intervention. Third, this Key-Stage information is available for previous cohorts of students, allowing us to test for balance in outcomes for prior cohorts and control for school level value added in difference-in-differences specifications. Finally, no additional testing was required to assess the impact of this program, thus the tests are not tailored to the intervention. Indeed, it has been shown that performance in these national age-11 exams is a strong predictor of later outcomes, including wages (DfE, 2013). This means we can estimate effects of the program on an outcome measure which has known benefits.

3.2 Recruitment

The target population for this study are state primary schools in England with above average Free School Meal eligibility (FSM) (which stood at 19 percent at the time of randomisation in 2013 (DfE, 2016)), and two or fewer classes per cohort.

The project developers were asked to recruit such primary schools in three regions in England in which they had capacity to deliver the program. The regions were the South West, East Midlands and North West. Each region contains a number of Local Authorities (LAs) that are responsible for the running of schools in that area.[8] In order to recruit schools the developers first had to obtain the approval of the relevant LAs. In the end, we recruited schools from 18 LAs (see Appendix 1 for the complete list). The aim of the recruitment was to eventually

[8]These are considerably larger than school districts in America with 152 currently operating in England. Unlike American school districts they have no power to raise finances to pay for school facilities; funding for education is provided to LAs from the central government who then allocate it across schools.

have 160 schools participate in the study. This total was determined by baseline power calculations (see Appendix Figure A.1).

Ultimately, 182 schools agreed to participate in the trial by sending back signed expression of interests. One of these schools was ineligible, as it would not have a cohort of students taking the age-11 tests during the evaluation period (it was a new school and only had younger year groups) and therefore was excluded. This left 181 schools that were to be randomized into treatment or control status as described in the randomization section below. All of these schools signed an agreement to grant us access to their NPD data prior to randomization. After randomization, the 89 schools selected for treatment additionally signed a Memorandum of Understanding which stated the responsibilities of the schools, practitioners, and the evaluation team.[9] These schools chose teachers who were to be involved in the program as long as they were teaching in academic years 4 and 5. The recruitment phase led to 6,436 participating students in the first cohort and 6,298 in the second cohort, for which we have administrative age-11 outcomes available.[10]

3.3 Representativeness

Figure 2 shows the geographical position of the schools in our sample, the red crosses denote schools of the treatment group and the blue crosses of the control group. We can see the schools come from three regions with the exception of one school in the south east of England. Table 1 shows how the schools within our sample compare with all schools nationwide and within the participating authorities, using information from students who completed their age-11 tests in 2011, three years prior to the intervention. In line with the recruitment strategy, pupils in our sample are slightly more likely to have Free School Meals (FSM) (22 percent) than pupils nationally (18 percent) or within their LA (19 percent). The students are more likely to possess a statement of Special Educational Needs (16 percent) than pupils nationally (14 percent) or locally (14 percent). As may be expected the average attainment at age-7 in these schools is lower than schools nationally

[9]In order to motivate schools to participate in this teacher development program we had to ensure that they did not perceive this intervention as useful for teacher assessment. One implication of this is that we could not collect and merge-in teacher-level information.

[10]There are 362 students (5 percent) for which the full set of demographics and attainment data was not available. This was approximately evenly split between treatment (172) and control groups (190).

(11 percent of a standard deviation). For the outcomes, age-11 tests, the students perform slightly worse in English (8 percent of a standard deviation), but achieve comparably to schools nationally or locally in maths (3 percent of a standard deviation lower). The proportion female and the cohort size are similar among our sample and schools locally and nationally. Taken as a whole, the schools in our sample contain slightly more disadvantaged students than an average school, and have a better value added in maths, but they are not distinctly different and therefore we have confidence in the external validity of the trial.

[Figure 2 goes here] [Table 1 goes here]

3.4 Randomization and compliance

We performed a pairwise stratified randomization of schools by LA with the aim of balancing the randomization at LA level (i.e. the pairing of schools for randomization was conducted within each LA). This was to ensure there were equal numbers of treated and control schools within each region and that they would be balanced in terms of unobservable local characteristics.

In order to pair similar schools within LAs we computed an index score using principal component analysis based on school level characteristics. These characteristics were taken from before the intervention in 2011, and consisted of the average maths and reading levels of students in their age-11 tests and the share of students eligible for FSM. Panel A of Table 2 shows the mean values of these variables and the index score of the sample, and the treatment and control groups.

Given the power calculations the evaluation had funding to implement the program in 80 schools and therefore the developers we asked to recruit at least 160 schools. Ultimately 182 expressed interest, of which 181 were eligible. There were not the funds to commit to funding the program in half of these schools, therefore treatment status was initally only allocated to schools for which we could construct an index score (8 schools had no age-11 test scores in 2011) and schools that did not operate as part of pair-franchise (6 schools). This left 167 schools of which 83 schools were assigned to treatment and 84 were assigned to control.[11]

[11]The randomization procedure is explained in more detail in Murphy et al. (2017)

When the 83 selected schools were informed that they would be treated, 16 no longer wished to take part, leaving 67 treatment schools.[12] The 14 previously excluded schools were then randomized into treatment and control groups. Pairs were randomly generated within reason for initial exclusion. For schools in pair-franchises, they were randomised as a pair, so that both schools were allocated to the same treatment status (two were assigned to treatment and four to control). Ultimately this resulted in 92 schools being allocated to control status and 89 allocated to treatment status, of which 73 initally participated in the program. Figure 3 presents the consort diagram, which traces the sample from recruitment, ramdomization to participation in the trial. During the course of the two year program five schools dropped out during the first year and four during the second year.[13] Meaning that 64 schools of the 89 schools assigned to treatment actually went through the full two-year intervention.

We examine dropouts in Appendix Tables A.1 and find evidence that dropouts were significantly different than the remaining sample for some characteristics, although no consistent picture emerges comparing the signficant characteristics from the first and second cohorts. We present summary statistics for all allocated schools, all schools that did not drop out, and for all dropout schools. Columns four and five report the raw differences and diffences conditional on the pair fixed effecuts used for the randomisation, for the first (Panel A) and second (Panel B) cohorts. For the first cohort schools that dropped out are larger and have students with slightly higher average age-7 attainment. However, for the second cohort these characteristics show no differences, with only the share of males being different.

In addition to schools being assigned to treatment and not being treated, students could also be assigned to treatment (by being enrolled in a treated school) but not treated. Individual-level treatment can differ from school-level treatment for two reasons. First, because they are in a class that is lead by a non-observed teacher. This occurs when there are two classes per cohort; the program only involves three teachers and therefore one class over the two cohorts would be left untreated. The NPD data does not allow us to determine how many teachers are

[12]Of the 16 schools not accepting treatment, 8 provided no reason, 5 reported staffing issues, one school change of school priorities, one due to school inspection, and one stating that that they only had 2 percent FSM and so should not be included

[13]Three of these schools this was due to teacher turnover, two due to having a new headteacher, two provided no reason, and two due to having to prioitise OfSted inspections

in a school year, but there is indicative evidence that this is the case - the proportion of a cohort being treated only falls below 50 percent in treated schools that participated in the study when the cohort size was above 34. Secondly, some students joined the school during the final year of primary school, meaning they take the age-11 tests with the treated cohort, but were not exposed to a program teacher since the treatment would have occurred before they joined. Therefore, the students within a year group that receive treatment might be non-random.

To determine if these excluded classes or new students are systematically different to the treated classes Appendix Table A.2 presents the characteristics of treated and non-treated students within treated schools. Here, we make use of the fact that all treatment schools that did not drop out provided us with lists of students that were taught by teachers in the program. Again there are some significant differences between treated and untreated students, but these differences are not consistent over cohorts. In the first cohort non-treated students have slightly lower age-7 test scores, are less likely to receive free school meals and are more likely to be male. In the second cohort there are no signifcant differences.[14]

As there are some significant observable (and potentially unobservable) differences between the those that were ultimately treated and those who were assigned to treatment (both at the school and student level) and these differences could be correlated with the size of the effect, our main conclusions will be based on intention to treat rather than realized participation. We also present Local Average Treatment Effects (LATEs) results for those schools and students who were actually treated, instrumenting with the assignment status.

[Figure 3 goes here]

3.5 Implementaton and fidelity

A full process evaluation took place alongside this quantitative study, including observation of the teacher training, interviews with staff involved in the treatment, and analysis of data on control schools' use of peer observation approaches. This

[14] As is expected untreated students come from schools that are signifcantly larger than treated students, because these schools will have a two class entry. However, there is no significant difference in school size when conditioning on pair fixed effects

qualitative evaluation was based on visits to 10 schools in 2 of the 3 implementation regions, to interview 19 staff and senior managers involved in the implementation. Follow up interviews were also conducted by telephone and email with 5 expert teachers in 5 schools, and information on progress provided by 4 other schools. Thus, we can report on the implementation of the peer-to-peer observation program in schools.

Many of the participating teachers reported having had some experience of using classroom observation in the past, for appraisal or development purposes. However these experiences were typically shorter (e.g. a 10 minute observation), more informal and less structured. For example, in describing a previous experience, one school pointed out that the process as a whole was not sufficiently structured to identify areas of improvement with sufficient accuracy and detail. Indeed, the structured nature of the program, particularly the requirement for record-keeping, was described as new, though teachers also believed it to be important to maintain rigor.

In general, fidelity was high, and schools were found to be implementing the peer-to-peer observation program according to the project design. The intensive 5 day training program may have been responsible for this high fidelity and indeed teachers rated this training highly, referring to it as outstanding' or high quality'. Many teachers reported that they felt prepared for the program from the outset as the training was well structured, interesting and based on evidence. The importance of teachers observing and not intervening was emphasized particularly strongly during the training, and teachers fully understood the reason for this rule and reported that they followed it.

The process evaluation concluded that the teachers viewed the program positively after implementing it. They reported finding certain features of the approach useful for their own practice, which reflect the potential mechanisms discussed earlier. First, they found it useful to reflect on their teaching and learning practice and welcomed the opportunity and 'space' within the timetable to reflect on their own practice. Second, they welcomed the input from peer observation, particularly with its emphasis on support, rather than performance management. For example, one teacher commented that the approach made it possible to convey to an underperforming teacher what they need to do to improve in a more supportive way. Teachers in particular reported positively on the experience of sharing practice

with teacher colleagues, shared planning, and identifying complementary skills.

4 Empirical approach

Prior to conducting the RCT we pre-committed to a set of specficiations and outcome measures in a Statistical Analysis Plan (SAP)[15], which was written three months before the beginning of the trial. The purpose of the SAP is to minimize conscious or sub-conscious decisions being made on the basis of results seen. The SAP contains details of the study design, sample size, randomization, chosen outcome measures, methodology and analysis plan, subgroup analysis. We now follow exactly the evaluation strategy that we set out initially and indicate the very few cases where we deviate.

Our primary analysis is conducted on an intention-to-treat' (ITT) basis. Specifically, we build up to from a univariate specification, only controlling for school assignment to treatment D_s, to the following model:

$$Y_{ips} = \alpha + \beta D_s + X_{it}{}'\delta + \pi_p + \varepsilon_{ips} \tag{1}$$

where the dependent variable Y_{ips} is the pupil i age-11 test score, in school pair p from school s. These students took their age-11 tests in the academic years 2014/15 (cohort 1) and 2015/16 (cohort 2). Students from the first cohort are only taught by teachers trained in the program for one year, whereas students from the second cohort are taught for two and teachers will be more accustomed to the system in the second year. To account for these differences the model is estimated for each cohort separately. β is our main parameter of interest and reflects the mean difference between those assigned to treatment and control groups. With successful randomization, a direct comparison of the means should be sufficient for determining the effect size. To improve the efficiency of the estimations we include X_{is} a vector of pupil characteristics. These are the student's average age-7 test scores (across maths and reading), and indicators for gender, special educational needs, English as a second language, ethnic minority status and FSM status. Given the pair-wise randomization structure, here we also include pair-fixed effects.

[15]This can be found at https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/lesson-study/

Throughout the analysis all standard errors are clustered at the school level.[16]

As noted previously, some schools that were assigned to treatment dropped out of the program. We therefore estimate LATEs estimated via two-stage least squares, where initial treatment allocation is used as an instrument for actual receipt of the intervention. It thereby corrects' the ITT estimate, by accounting for the non-compliance of some schools or students. The actual receipt of the intervention is defined in two ways. First, at the school cohort level (T_s), where we define a school to be treated if we received confirmation from the school at the end of each academic year that they participated. Second, at the student level (T_{is}), if we received confirmation from the school that the student was taught by an observed teacher.

$$T_{is} = \alpha + \beta_1 D_s + X_i'\delta + \pi_p + \varepsilon_{ips} \tag{2}$$

$$Y_{ips} = \alpha + \beta_2 \hat{T}_{is} + X_i'\delta_2 + \pi_p + \tau_{ips} \tag{3}$$

In an alternate specification, we exploit the panel nature of the administrative data, which increases our sample size dramatically, and allows us to perform a difference-in-differences analysis. Here we introduce the subscript t to the dependent variable Y_{ipst} and school assignment D_{st}. In addition to pair fixed effects we include a set of year effects (μ_t). The difference-in-differences analysis includes all years from 2008/9 up to the start of the trial (2012/13) as control years and the corresponding treatment year only. This means for the first cohort we omit 2013/14 from the analysis and we omit 2014/15 for the analysis of the second cohort.

$$Y_{ipst} = \alpha + \beta D_{st} + X_{it}'\delta + \pi_p + \mu_t + \varepsilon_{ipst} \tag{4}$$

[16]For the main results in Table 3 we provide simulated Fisher exact p-values (see also Appendix Figures A.2 and A.3). Due to the large sample size of this trial, these are very similar.

5 Results

5.1 Balance at baseline

Before presenting the effects of peer-to-peer observation and feedback on student outcomes, Table 2 shows summary baseline statistics for the treatment and control schools, both at the school-level (unweighted) in a pre-treatment year 2011 (Panel A) and at the student level for the treated cohort 1 (Panel B) and cohort 2 (Panel C). The school-level information shown in Panel A shows age-11 outcomes in maths, english, and the share of free school meal recipients. This is the school-level information that was available pre-intervention and that we combined into an Index Score to perform the pairwise matching as described above in section 3.4. Panel A confirms that the randomisation generated balance on the school characteristics that were used to generate the pairs for randomisation. Panels B and C of Table 2 present the balancing on a wider range of student characteristics for the cohorts used in the analysis cohorts. All cells in columns 4 and 5 show similar student characteristics and attainments for treatment and control groups. Therefore, we conclude that the randomisation based on pairwise randomisation dependent on school characteristics from previous cohorts generated balanced treatment and control groups.

[Table 2 goes here]

5.2 Effects on pupil attainment

5.2.1 Cross-sectional results

Table 3 presents unconditional estimates of program assignment on the pre-specified primary outcome, the combined test score (average maths and reading scores), as well as secondary outcomes, maths, reading, SPAG, and science. Combined age-11 national test percentile for the first treated cohort is 47.25 and 46.13 for the first control cohort, though this difference is not statistically significant at conventional levels. For the second cohort (which was treated for two years) there is also no significant difference between the treatment and control groups with the respective

average percentile scores being 45.50 and 45.55. A similar pattern is found for the secondary outcomes all of which the treatment group is not statistically significant different from the control group for either cohort. Note that all these differences are of assigned treatment and control groups and so represent unconditional ITT effects. Column 5 reports simulated Fisher p-values for these effects which in no instance are close to rejecting the null hypothesis of a zero effect.

[Table 3 goes here]

5.2.2 Main results

The main results are presented in Table 4, where we continue to report ITT estimates using increasingly relaxed specifications. Column 1 is identical to the raw cross sectional results shown in column 3 of Table 3, where the coefficients represent the impact on national percentile rank from the school being assigned to treatment. Moving through columns 2-4 we subsequently add pair-wise fixed effects, prior student age-7 test scores, and student demographics. The inclusion of these additional controls do not result in any significant change in any of the estimates. This is not surprising given the strong balancing with respect to these observable characteristics reported above. However, their inclusion does reduce the size of the clustered standard errors, in particular the inclusion of the pair fixed effects dramatically increases the precision of our estimates. In column 5 we report standardised effect sizes, rather than impact on national percentile rank. For our main test score outcome we can reject effect sizes of 8.7 percent of a standard deviation for cohort 1 and of 11 percent of a standard deviation for cohort 2, based on two-sided confidence intervals at the 95 percent level of statistical significance.

[Table 4 goes here]

5.2.3 IV analysis

In Table 5 we turn to the LATE estimates to estimate the impact of the program on those that actually went through with it, as the null results found in the previous table could be due to schools not participating in the program.

[Table 5 goes here]

For comparison the first column repeats the estimates from the most relaxed specification from Table 4, in which we control for student characteristics, prior attainment and pair fixed effects. Column 2 presents estimates from instrumenting the school-cohort-level treatment with the school-level random assignment. For both cohorts, the estimated LATEs are larger in magnitude but remain insignificant. For example, the ITT effect on the combined age-11 test pecentile was 0.27 for cohort 1 and the LATE is now estimated at 0.335. We can gauge the scaling by noting the size of the first stage coefficient of about 0.8, implying that 80 percent of schools assigned to the treatment went through with it.

Column 4 presents results from the student-level LATE analysis. Here, we instrumenting individual level treatment with the school-level random assignment. Note that the corresponding first stage estimates of the student-level IV are smaller compared to the school-level IV, now estimated at 0.667 and 0.674 for cohorts 1 and 2. This is precisely because not all students in an assigned treatment cohort were treated. Taking the ratio of the two first stages we can see that 82 percent of students in schools that went through with the treatment were taught by a trained teacher Turning to student test scores, the estimates remain insignificant. The LATE of our primary outcome measure now stands at 0.405 with a standard error of 1.482 in column 4 Panel A, for example.

Columns 3 and 5 provide the corresponding standardised estimates. For cohort 1 (Panel A) the standardised ITT effect is 0.011 (from column 5 of Table 4) and the LATEs are now estimated at 0.013 for school treatment (Table 5 column 3) and 0.016 for student treatment (Table 5 column 5). For the second cohort, which would have experienced two years of the treatment, the standardised effects are approximate doubled in size (at 0.023, 0.029 and 0.035 respectively) but remain close to zero and indistinguishable from zero at conventional levels of statistical significance.

To remain in accordance with the SAP we estimate the equivalent table using the difference-in-differences specification This will be using the variation within treated schools across time and between treated and untreated schools. These estimates are presented in Appendix Table A.3 , both the estimates from the ITT and both sets of instrumented specifications change very little when additionally

using these pre-treatment cohorts, which is reassuring given the randomised nature of the experiment. However, we also find that the inclusion of the additional cohorts does not reduce the standard errors, which in some cases are larger. This provides an potentially interesting finding that including additional cohorts does not improve the power of a RCT when the standard errors are clustered at the time invariant unit of treatment.

5.2.4 Heterogeneity

Again to remain in accordance with the SAP, we finally present sub-group ITT analysis in Table 6. The five subgroups analysed are students who are eligible for free school meals (FSME), speak English as additional language (ESL), belong to an ethnic minority, are low achievers in terms of their age-7 outcomes, or are male.[17]

[Table 6 goes here]

Out of the forty interaction terms estimated here, three are statistically significant at the five percent level (or higher). In cohort 2 only, the overall effect as well as the effect on maths scores for girls is not statistically significantly different from zero but boys seem to be negatively affected compared to girls. In contrast, in cohort 1 the interactions for minority and low age-7 test scores are positive and statistically significant for the SPAG outcome. Given the inconsistent pattern and the fact that the analysis presented in columns 3 to 8 of Table 6 is not part of the pre-registered analysis plan, we conclude that there is little evidence for significant heterogeneity.

6 Discussion

To what extent do our results differ from Taylor and Tyler (2012), who find that teacher observation and feedback can lead to improved test scores in Cincinnati? The Cincinnati study estimates effects on maths test score outcomes only, so in order to best compare the two studies we should consider test score outcomes in

[17] Appendix Table A.4 shows corresponding total effect sizes.

maths only. The Cincinnati study finds positive effects of 11.2 percent of a standard deviation in the first year after the (short) intervention, and effects of 15.8 percent of a standard deviation two years after the intervention (and even larger effects later on).[18] Applied to our setting, this first estimate can be best compared to our estimates from Table 4 for maths outcomes for cohort 1, where the intervention started almost two years before students took the test, and the second to our cohort 2, where students started being treated almost three years after the start of the intervention (both had one year of being taught by an un-observed teacher before the examination). For both cohorts, the 95-percent confidence intervals do not include the point estimates from Taylor and Tyler (2012). We can reject effects of up to 11.04 (cohort 1) and 12.62 percent of a standard deviation (cohort 2) respectively, based on a two-sided test of null effects.[19] Moreover, the Cincinnati study presents in the main results table even larger effects of about twenty percent of a standard deviation for academically weaker students. In contrast, we find statistically insignificant and similar or even marginally smaller effect sizes for maths outcomes for students with previously low-age 7 tests in our heterogeneity analysis in column 4 of Appendix Table A.4. These estimates can reject the effect sizes found by Taylor and Tyler (2012) at the 99 percent level of statistical significance.

Our finding that teacher peer observation carried out in this structured manner does not lead to improved pupil performance thus highlights a clear difference between our results and the most well-known study of teacher observation and its impact on pupil achievement. There are three potential reasons why this may be the case.

First, and most notably, Cincinnati teachers were formally scored, with the results carrying explicit consequences, including impact on promotions and tenure, and potential non-renewal of the teacher's contract. Our trial did not involve such consequences and was designed purely to improve teacher performance through discussion and feedback.

Secondly, the Cincinnati study involved filming of observed classes and being observed by non-peer experts. This may have had an unintended effect of encouraging students to behave differently when being filmed and observed (e.g. they

[18] see Table 5 of Taylor and Tyler (2012)

[19] This is using the estimates for maths outcomes shown in Table 4 column 5 and the 95 percent level of statistical significance.

may have been better behaved), which may have led to improved test scores. It may have also resulted in more accuracy in the observation process in Cincinnati with teachers able to refresh their memory of what they observed after the fact. Moreover having an external expert there to provide feedback may have resulted in the delivery of more informative, accurate and frank feedback to teachers.

Thirdly, and finally, our program took place under experimental conditions. Our study therefore overcomes issues typical of quasi-experimental studies. For example, in our study, there is no difference in the characteristics of treated versus untreated teachers; in the Cincinnati setup, younger teachers were evaluated first, which could result in upward bias of the results if younger teachers have higher growth in value-added than older teachers[20]

We believe these key differences between the studies suggest that teacher peer-to-peer observation may be not be effective unless coupled with incentives and external evaluators. As such, besides providing convincing evidence on the efficacy of teacher observation programs, this paper also brings new evidence on the potential mechanisms through which the Cincinnati study may have generated a positive result.

While we can reject the effect sizes of the Cincinnati Study, smaller positive effect sizes cannot be rejected. Chetty et al. (2014) estimate that replacing the bottom 5% of teachers in terms of value added with an average teacher would increase the present value of students' lifetime income by about $250,000 per classroom. How do our estimates compare to teacher effectiveness at the bottom end? We reject the effect sizes of one-standard deviation better teachers in terms of value-added presented in Hanushek and Rivkin (2010). Therefore, teacher peer-to-peer observation and feedback studied here is unfit to close the gap between ineffective and effective teachers, or indeed between ineffective and average teachers. Thus, while the evidence presented in this trial does not rule out that this type of teacher observation and feeback may have small positive effects on student performance, it clearly cannot solve the policy-maker's problem of huge variation in teacher effectiveness.

[20]There is also evidence that the teachers in the Cincinnati sample were somehow able to affect the timing of their observation; when using scheduled rather than actual interview date the effect sizes are halved. Such manipulation is not possible in our setup. Our setup also minimizes the possibility of attrition (as the compulsory test scores are collected centrally).

7 Conclusions

The unpredictability of teacher productivity has led to growing efforts to measure it in the classroom. Teacher peer observation is an obvious means of doing so and many schools adopt such practices, either as a means to identify good teachers, or to improve their existing labor force. By implementing a large-scale randomized control trial across primary schools in England, we attempt to provide robust evidence on the efficacy of teacher peer observation as a teacher development tool. Our results - that we find no positive impact of teacher observation on pupil performance in reading, maths, science or grammar, across any subgroup - are in contrast to the limited body of research in this area, which has pointed to a positive role for teacher peer observation on pupil outcomes.

Our study had high fidelity and its large-scale nature means we are able to rule out effect sizes larger than 8.7 percent of a standard deviation in national age-11 test scores for our first cohort, and larger than 11 percent of a standard deviation for our second cohort.

However, our study does have a number of limitations which should be noted. First, our outcome measures age-11 scores in reading and maths are obtained a year after the end of implementation of the intervention (and almost two and three years after its start) and therefore represent a medium-term outcome. This may have reduced the possibility of finding an effect, though, as noted above, it is precisely where one would hope to find it. However, we cannot say what the impact might have been directly after the implementation, or indeed in the long-term, for example as there could have been incremental changes to teacher practice. Additionally, our outcome measures are purely academic and we therefore cannot say whether peer-to-peer observation and feedback may have had an impact on non-cognitive pupil outcomes such as well-being or emotional development, though the program was never intended to affect such outcomes.

A final caveat concerns our finding that many schools already implement some form of peer-to-peer feedback, albeit in a less structured and comprehensive way and with lesser intensity than in this intervention. Our research cannot quantify what the impact of the two-year peer-to-peer observation and feedback intervention would be compared to schools who do not carry out any individual activities, rather it is a comparison to business as usual. Similarly, we cannot conclude that any of

the component parts have no impact, given that schools in the control group may also perform many of them in less structured ways.

Our results are likely generalisable since they are based on a large sample of primary schools. The use of teacher observation and feedback is widespread and gaining traction and there are many commonalities in approaches used across schools in the UK and internationally. We believe that the results of this research are highly relevant for schools carrying out these activities. Moreover, as described above, our results indicate that teacher observation and feedback is not effective in the absence of teacher incentives and non-peer feedback and cannot be used to significantly reduce differences in teacher effectiveness.

References

Aaronson, D., L. Barrow, and W. Sander (2007). Teachers and student achievement in the chicago public high schools. *Journal of labor Economics 25*(1), 95–135.

Angrist, J. D. and V. Lavy (2001). Does teacher training affect pupil learning? evidence from matched comparisons in jerusalem public schools. *Journal of labor economics 19*(2), 343–369.

Chetty, R., J. N. Friedman, and J. E. Rockoff (2014). Measuring the impacts of teachers ii: Teacher value-added and student outcomes in adulthood. *American Economic Review 102*(5), 1805–1831.

DfE (2013). Reading and maths skills at age 10 and earnings in later life: a brief analysis using the british cohort study. Technical report, Department for Education.

DfE (2016). *Schools, Pupils and Their Characteristics, January 2016*. Dandy Booksellers Limited.

Dixit, A. (2002). Incentives and organizations in the public sector: An interpretative review. *Journal of human resources*, 696–727.

Fernandez, C., J. Cannon, and S. Chokshi (2003). A us–japan lesson study collaboration reveals critical lenses for examining practice. *Teaching and teacher education 19*(2), 171–185.

Garet, M. S., A. J. Wayne, F. Stancavage, J. Taylor, M. Eaton, K. Walters, M. Song, S. Brown, S. Hurlburt, P. Zhu, et al. (2011). Middle school mathematics professional development impact study: Findings after the second year of implementation. ncee 2011-4024. *National Center for Education Evaluation and Regional Assistance*.

Garet, M. S., A. J. Wayne, F. Stancavage, J. Taylor, K. Walters, M. Song, S. Brown, S. Hurlburt, P. Zhu, S. Sepanik, et al. (2010). Middle school mathematics professional development impact study: Findings after the first year of implementation. ncee 2010-4009. *National Center for Education Evaluation and Regional Assistance*.

Goodman, S. and L. Turner (2010). Teacher incentive pay and educational outcomes: Evidence from the nyc bonus program. program on education policy and governance working papers series. pepg 10-07. *Program on Education Policy and Governance, Harvard University*.

Hanushek, E. A. and S. G. Rivkin (2010). Generalizations about using value-added measures of teacher quality. *American Economic Review 100*(2), 267–71.

Harris, D. N. and T. R. Sass (2011). Teacher training, teacher quality and student achievement. *Journal of public economics 95*(7-8), 798–812.

Jackson, C. K. and E. Bruegmann (2009). Teaching students and teaching each other: The importance of peer learning for teachers. *American Economic Journal: Applied Economics 1*(4), 85–108.

Jacob, B., J. E. Rockoff, E. S. Taylor, B. Lindy, and R. Rosen (2016). Teacher applicant hiring and teacher performance: Evidence from dc public schools. Technical report, National Bureau of Economic Research.

Jacob, B. A. and L. Lefgren (2008). Can principals identify effective teachers? evidence on subjective performance evaluation in education. *Journal of labor Economics 26*(1), 101–136.

Kane, T. J., J. E. Rockoff, and D. O. Staiger (2008). What does certification tell us about teacher effectiveness? evidence from new york city. *Economics of Education review 27*(6), 615–631.

Lavy, V. (2009). Performance pay and teachers' effort, productivity, and grading ethics. *American Economic Review 99*(5), 1979–2011.

Muralidharan, K. and V. Sundararaman (2011). Teacher performance pay: Experimental evidence from india. *Journal of political Economy 119*(1), 39–77.

Murphy, R., F. Weinhardt, and G. Wyness (2017). Lesson study evaluation report and executive summary.

Neal, D. (2011). The design of performance pay in education. In *Handbook of the Economics of Education*, Volume 4, pp. 495–550. Elsevier.

Papay, J., J. Tyler, and E. Taylor (2018). Using teacher evaluation data to drive instructional improvement: Evidence from the evaluation paternatship program in tennessee (2015-2020). Unpublished.

Perry, R. R. and C. C. Lewis (2009). What is successful adaptation of lesson study in the us? *Journal of Educational Change 10*(4), 365–391.

Rivkin, S. G., E. A. Hanushek, and J. F. Kain (2005). Teachers, schools, and academic achievement. *Econometrica 73*(2), 417–458.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review 94*(2), 247–252.

Rothstein, J. (2015). Teacher quality policy when supply matters. *American Economic Review 105*(1), 100–130.

Springer, M. G., D. Ballou, L. Hamilton, V.-N. Le, J. Lockwood, D. F. McCaffrey, M. Pepper, and B. M. Stecher (2011). Teacher pay for performance: Experimental evidence from the project on incentives in teaching (point). *Society for Research on Educational Effectiveness*.

Steinberg, M. P. and L. Sartain (2015). Does teacher evaluation improve school performance? experimental evidence from chicago's excellence in teaching project. *Education Finance and Policy 10*(4), 535–572.

Taylor, E. S. and J. H. Tyler (2012). The effect of evaluation on teacher performance. *American Economic Review 102*(7), 3628–51.

Weisberg, D., S. Sexton, J. Mulhern, D. Keeling, J. Schunck, A. Palcisco, and K. Morgan (2009). The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness. *New Teacher Project*.

Tables and figures

Figure 1: Timeline of intervention

Calendar Year / School Year	2010/2011	2011/2012	2012/2013	2013/2014	2014/2015	2015/2016
Year 2 (Age-7 tests - Controls)	**Cohort 1**	**Cohort 2**	Cohort 3	Cohort 4	Cohort 5	Cohort 6
Year 3	Cohort 0	**Cohort 1**	**Cohort 2**	Cohort 3	Cohort 4	Cohort 5
Year 4	Cohort -1	Cohort 0	**Cohort 1**	**Cohort 2**	Cohort 3	Cohort 4
Year 5	Cohort -2	Cohort -1	Cohort 0	**Cohort 1**	**Cohort 2**	Cohort 3
Year 6 (Age-11 tests - Outcomes)	Cohort -3	Cohort -2	Cohort -1	Cohort 0	**Cohort 1**	**Cohort 2**

Notes: Red square shows treatment period and cohorts.

Figure 2: Treatment and control schools

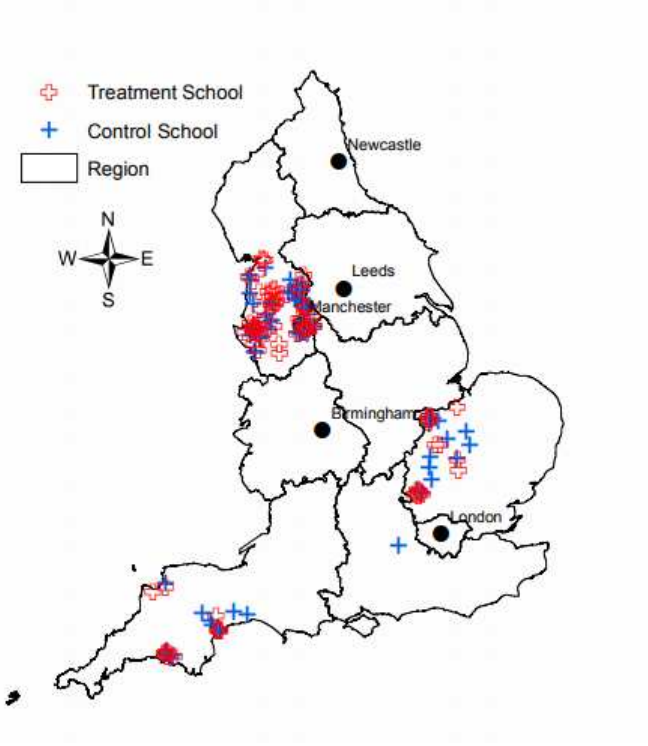


Figure 3: Consort flow diagram

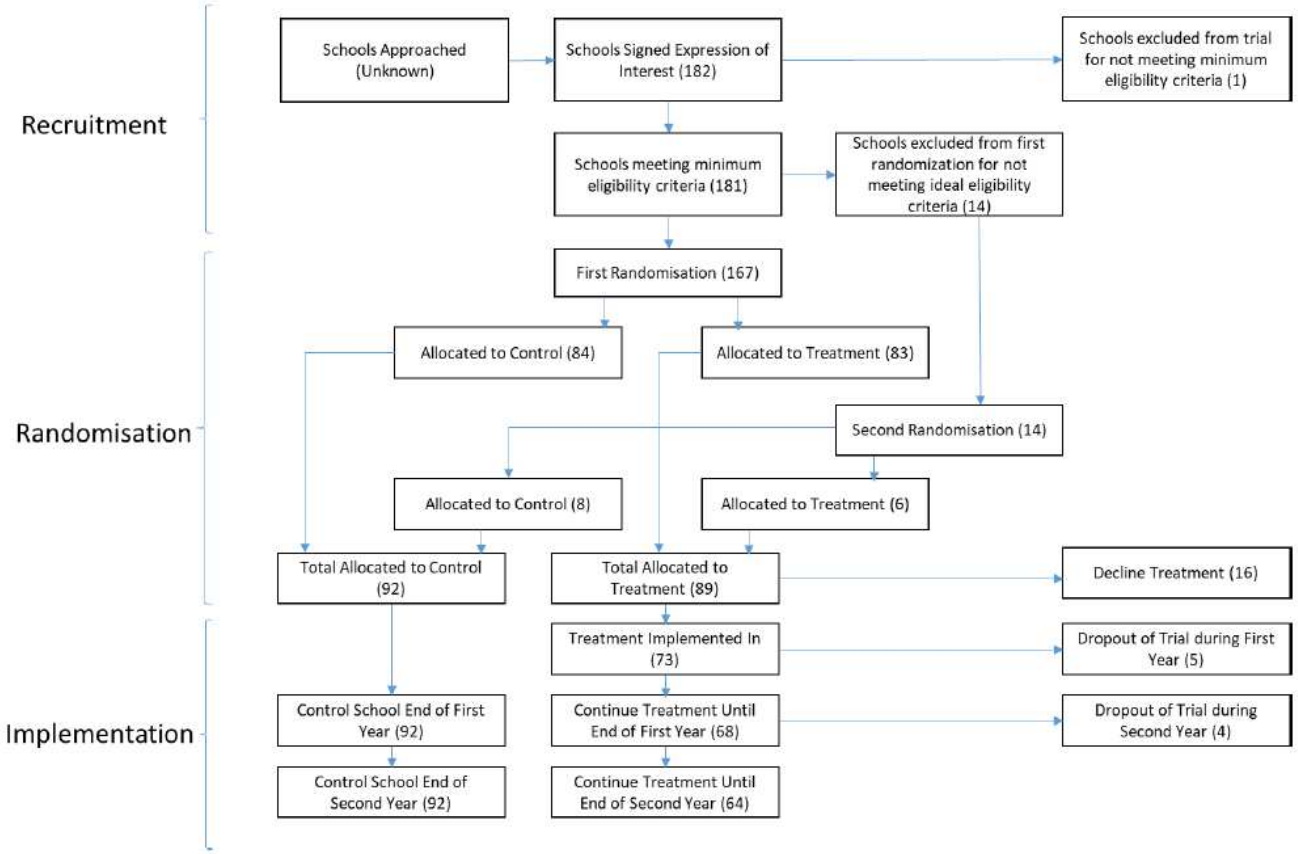


Table 1: National and local representativeness of sample

Variable	(1) National	(2) Local	(3) Sample	(4) (1)-(2)	(5) (2)-(3)
Age-7 Test	15.709	15.643	15.269	-0.445	-0.412
	[3.917]	[3.897]	[3.787]	(0.117)	(0.123)
Age-11 Maths Level	3.047	3.036	3.015	-0.033	-0.024
	[0.986]	[0.990]	[0.944]	(0.030)	(0.033)
Age-11 English Level	2.988	2.970	2.909	-0.079	-0.067
	[0.974]	[0.981]	[0.953]	(0.032)	(0.034)
Share Free School Meals	0.181	0.192	0.223	0.042	0.034
	[0.385]	[0.394]	[0.416]	(0.012)	(0.013)
Share Female	0.489	0.492	0.499	0.010	0.008
	[0.500]	[0.500]	[0.500]	(0.006)	(0.006)
Share Special Edu. Needs	0.137	0.142	0.160	0.023	0.020
	[0.344]	[0.349]	[0.367]	(0.008)	(0.008)
School Size	51.349	45.757	48.712	-2.667	3.254
	[29.472]	[26.838]	[23.438]	(2.444)	(2.626)
Students in 2011/age-11 cohort	554,768	69,346	6,372		

Notes: This table shows baseline characteristics for a pre-treatment cohort sitting the age-11 tests in maths and english in 2011. Note that for this cohort age-11 test scores were only avaialable to us in levels at the time of the randomsation so that these are not percentalised. Column 1 includes all students of that cohort, column 2 only students in the same Local Authority and column 3 students of the schools that were part of the trial. Standard deviations of variables shown in square parenthesis. Standard errors clustered at the school level shown in round parenthesis.

Table 2: Randomisation tests: pre-period, cohort 1 and cohort 2

	(1) Sample	(2) Control	(3) Treated	(4) (2)-(3)	(5) (2)-(3)
Panel A: School-level 2011					
Age-11 Maths Level	3.012	3.018	3.007	-0.011	-0.016
	[0.328]	[0.313]	[0.344]	(0.051)	(0.026)
Age-11 English Level	2.91	2.917	2.902	-0.015	-0.015
	[0.363]	[0.350]	[0.378]	(0.056)	(0.03)
Share Free School Meals	0.244	0.24	0.248	0.008	0.005
	[0.168]	[0.172]	[0.164]	(0.026)	(0.018)
Index Score	-0.018	0.016	-0.052	-0.067	-0.071
	[1.433]	[1.375]	[1.497]	(0.223)	(0.077)
Panel B: Cohort 1					
Age-7 Test	15.540	15.614	15.469	0.145	0.166
	[3.566]	[3.539]	[3.591]	(0.221)	(0.114)
Free School Meals	0.236	0.237	0.236	0.001	0.012
	[0.425]	[0.425]	[0.425]	(0.025)	(0.014)
Special Edu. Needs	0.138	0.139	0.136	0.003	0.008
	[0.345]	[0.346]	[0.343]	(0.013)	(0.010)
Gender: Male	0.503	0.502	0.505	-0.003	-0.001
	[0.500]	[0.500]	[0.500]	(0.013)	(0.011)
Minority	0.218	0.240	0.196	0.044	0.037
	[0.413]	[0.427]	[0.397]	(0.054)	(0.024)
ESL	0.179	0.210	0.148	0.062	0.054
	[0.383]	[0.408]	[0.355]	(0.050)	(0.022)
School Size	47.681	46.896	48.434	-1.538	-0.293
	[25.941]	[24.641]	[27.113]	(5.879)	(2.450)
Panel C: Cohort 2					
Age-7 Test	15.935	15.823	16.050	-0.227	-0.150
	[3.444]	[3.455]	[3.430]	(0.216)	(0.127)
Free School Meal	0.233	0.240	0.227	0.013	0.014
	[0.423]	[0.427]	[0.419]	(0.025)	(0.014)
Special Edu. Need	0.134	0.126	0.141	-0.015	-0.010
	[0.340]	[0.332]	[0.348]	(0.016)	(0.011)
Gender: Male	0.507	0.504	0.510	-0.006	-0.007
	[0.500]	[0.500]	[0.500]	(0.012)	(0.009)
Minority	0.220	0.243	0.196	0.047	0.038
	[0.414]	[0.429]	[0.397]	(0.054)	(0.024)
ESL	0.183	0.217	0.149	0.068	0.049
	[0.387]	[0.412]	[0.356]	(0.050)	(0.023)
School Size	50.236	47.422	53.099	-5.678	-2.262
	[31.747]	[23.257]	[38.302]	(8.677)	(3.789)
Pair FX					X

Notes: Panel A is for sampled schools in pre-period. Panels B and C show balancing at the student level for cohorts 1 and 2. Number of obs. for sample in panels A/B/C: 167/6,436/6,298. Standard deviations of variables shown in square parenthesis in columns 1-3. Standard errors clustered at the school level shown in round parenthesis in columns 4-5.

Table 3: Cross-sectional results

	(1) Treatment	(2) Control	(3) Difference	(4) Standardised	(5) Fisher p-value
Panel A: Cohort 1					
Test Score	47.25	46.13	1.12	0.044	
	(0.46)	(0.44)	(1.734)	(0.068)	0.376
Maths	48.13	46.26	1.87	0.066	
	(0.51)	(0.49)	(1.87)	(0.067)	0.384
Reading	46.38	46.00	0.376	0.014	
	(0.49))	(0.48)	(1.72)	(0.062)	0.860
SPAG	48.15	47.17	0.98	0.035	
	(0.48)	(0.48)	(1.73)	(0.063)	0.625
Science	4.26	4.27	-0.01	-0.014	
	(0.11)	(0.11)	(0.04)	(0.058)	0.875
Panel B: Cohort 2					
Test Score	45.50	45.55	-0.05	-0.00	
	(0.45)	(0.46)	(1.66)	(0.07)	0.982
Maths	46.87	46.53	0.34	0.012	
	(0.49)	(0.51)	(1.81)	(0.07)	0.892
Reading	44.13	44.58	-0.45	-0.016	
	(0.49)	(0.50)	(1.72)	(0.06)	0.856
SPAG	45.35	45.35	-1.31	-0.047	
	(0.49)	(0.51)	(1.73)	(0.06)	0.566

Notes: This tables shows results of unconditional cross-sectional comparisions, separately for cohorts 1 and 2 (Specification 1 in main text), separately for cohorts 1 (Panel A) and cohort 2 (Panel B). Test Score refers to combined reading and maths tests at age 11. Science scores were only recorded for cohort 1. Number of observations: cohort 1 (cohort 2) 6,436 (6,298). Standard errors in parenthesis in column 3 are clustered at school level. Column 5 shows Fisher exact p-values for null effects, based on 10,000 simulations (see Appendix Figures A.2 and A.3.)

Table 4: Main results

	(1)	(2)	(3)	(4)	(5)
Panel A: Cohort 1					
Test Score	1.122	1.301	0.444	0.270	0.011
	(1.734)	(1.089)	(0.990)	(0.990)	(0.039)
Maths	1.867	2.064	1.186	0.897	0.032
	(1.872)	(1.240)	(1.128)	(1.130)	(0.040)
Reading	0.376	0.538	-0.299	-0.357	-0.013
	(1.718)	(1.054)	(0.981)	(0.962)	(0.035)
SPAG	0.979	1.229	0.335	-0.237	-0.009
	(1.732)	(1.222)	(1.155)	(1.115)	(0.040)
Science	-0.010	-0.018	-0.038	-0.040	-0.063
	(0.039)	(0.027)	(0.025)	(0.025)	(0.039)
Panel A: Cohort 2					
Test Score	-0.053	-0.004	0.767	0.597	0.023
	(1.660)	(1.226)	(1.145)	(1.132)	(0.045)
Maths	0.341	0.493	1.291	1.003	0.035
	(1.805)	(1.413)	(1.328)	(1.299)	(0.046)
Reading	-0.448	-0.502	0.244	0.192	0.007
	(1.719)	(1.162)	(1.095)	(1.099)	(0.040)
SPAG	-1.305	-0.925	-0.076	-0.585	-0.021
	(1.732)	(1.133)	(1.141)	(1.105)	(0.039)
Pair FX		X	X	X	X
Age-7 test score			X	X	X
Demographics				X	X
Standardised					X

Notes: This tables shows results of the intervention at age-11 on average english and maths test scores [Test Score (age-11)], maths test scores, reading test scores, scores for spelling, punctuation and grammar [SPAG] and Science, separately for cohort 1 [Panel A] and cohort 2 [Panel B]. Moving from left the right, additional variables are added as controls as indicated at the bottom of the table. Column (1) shows estimates of specification (1) in the text. Science scores were only recorded for cohort 1. Number of observations: cohort 1 (cohort 2) 6,436 (6,298). Standard errors clustered at the school level in parenthesis.

Table 5: IV analysis

	(1)	(2)	(3)	(4)	(5)
	ITT	School LATE		Student LATE	
Panel A: Cohort 1					
Test Score	0.270	0.335	0.013	0.405	0.016
	(0.989)	(1.226)	(0.048)	(1.482)	(0.058)
Maths	0.897	1.114	0.040	1.345	0.048
	(1.130)	(1.396)	(0.050)	(1.698)	(0.060)
Reading	-0.357	-0.444	-0.016	-0.536	-0.019
	(0.962)	(1.203)	(0.044)	(1.449)	(0.053)
SPAG	-0.237	-0.294	-0.011	-0.355	-0.013
	(1.115)	(1.388)	(0.050)	(1.673)	(0.061)
Science	-0.040	-0.050	-0.078	-0.061	-0.094
	(0.025)	(0.031)	(0.048)	(0.038)	(0.059)
First Stage		0.805		0.667	
		(0.033)		(0.032)	
Panel B: Cohort 2					
Test Score	0.597	0.747	0.029	0.886	0.035
	(1.132)	(1.406)	(0.055)	(1.679)	(0.066)
Maths	1.003	1.253	0.044	1.487	0.052
	(1.299)	(1.608)	(0.057)	(1.928)	(0.068)
Reading	0.192	0.240	0.009	0.285	0.010
	(1.099)	(1.372)	(0.049)	(1.630)	(0.059)
SPAG	-0.585	-0.731	-0.026	-0.868	-0.031
	(1.105)	(1.386)	(0.049)	(1.641)	(0.058)
First Stage		0.800		0.674	
		(0.035)		(0.033)	
Standardised			X		X

Notes: Column (1) is the ITT effect and indentical to column (4) of Table 4: Pair-FX, Age-7 test scores and student demographics are included as controls. Columns (2) and (3) show results when random assignment to the treatment is used as instrument for actual school-level takeup. Columns (4) and (5) repeat this exercise but using student-level information about take-up. Number of observations: cohort 1 (cohort 2) 6,436 (6,298). Standard errors clustered at the school level in parenthesis.

Table 6: Heterogeneity

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Test Scores		Maths		Reading		SPAG	
	Main Effect	Interaction	Main Effect	Interaction	Main Effect	Interaction	Main Effect	Interaction
Panel A: Cohort 1								
Share Free School Meals	0.023	-0.015	0.046	-0.007	-0.007	-0.016	0.031	-0.064
	(0.041)	(0.043)	(0.043)	(0.045)	(0.038)	(0.045)	(0.044)	(0.048)
ESL	0.011	-0.014	0.025	0.035	-0.005	-0.063	-0.031	0.135
	(0.043)	(0.067)	(0.044)	(0.069)	(0.038)	(0.068)	(0.044)	(0.080)
Minority	0.015	0.004	0.026	0.062	-0.000	-0.056	-0.032	0.164
	(0.042)	(0.095)	(0.043)	(0.063)	(0.037)	(0.061)	(0.042)	(0.071)
Low age-7 test score	0.013	0.038	0.039	0.032	-0.019	0.049	-0.007	0.116
	(0.043)	(0.055)	(0.044)	(0.058)	(0.039)	(0.054)	(0.046)	(0.057)
Gender: male	0.001	0.033	0.031	0.023	-0.033	0.042	-0.012	0.051
	(0.043)	(0.031)	(0.047)	(0.039)	(0.040)	(0.035)	(0.045)	(0.034)
Panel B: Cohort 2								
Share Free School Meals	0.035	-0.013	0.051	-0.004	0.017	-0.018	0.003	-0.014
	(0.048)	(0.045)	(0.050)	(0.046)	(0.042)	(0.048)	(0.044)	(0.047)
ESL	0.012	0.059	0.018	0.104	0.009	0.004	-0.034	0.098
	(0.048)	(0.074)	(0.049)	(0.080)	(0.042)	(0.069)	(0.043)	(0.087)
Minority	0.017	0.058	0.030	0.077	0.005	0.033	-0.025	0.069
	(0.051)	(0.067)	(0.052)	(0.068)	(0.045)	(0.069)	(0.045)	(0.070)
Low age-7 test score	0.032	-0.011	0.044	0.036	0.015	-0.022	-0.007	0.032
	(0.048)	(0.055)	(0.051)	(0.064)	(0.041)	(0.052)	(0.045)	(0.059)
Gender: male	0.063	-0.062	0.095	-0.088	0.021	-0.019	0.013	-0.031
	(0.046)	(0.036)	(0.049)	(0.040)	(0.043)	(0.039)	(0.042)	(0.035)

Notes: This tables shows estimates for main effects and interactions for age-11 outcomes in overall test scores (col 1-2) maths (col 3-4), reading (col 5-6) and spelling, punctuation and grammar (col 7-8). Columns 3-8 is not part of the pre-registered analysis plan of the intervention. All specfications include pair FX and age-7 test scores as controls. Estimates are standardized. Standard errors clustered at school level in parenthesis.

Appendix

Appendix 1: Participating Local Authorities

341	Liverpool
342	St Helens
343	Sefton
344	Wirral
352	Manchester
353	Oldham
354	Rochdale
356	Stockport
357	Tameside
821	Luton
823	Central Bedfordshire
867	Bracknell Forest
873	Cambridgeshire
874	Peterborough, City of
878	Devon
879	Plymouth, City of
888	Lancashire
896	Cheshire West and Chester

Tables and Figures

Table A.1: Analysis of School-Level Dropout

	(1) Sample	(2) Dropout	(3) Stayer	(4) (2)-(3)	(5) (2)-(3)
Panel A: Cohort 1					
Age-7 Test	15.540	15.455	15.553	-0.098	0.560
	[3.566]	[3.372]	[3.596]	(0.309)	(0.238)
Share Free School Meals	0.236	0.261	0.233	0.029	-0.002
	[0.425]	[0.439]	[0.423]	(0.031)	(0.030)
Gender: Male	0.503	0.514	0.502	0.013	0.019
	[0.500]	[0.500]	[0.500]	(0.017)	(0.026)
Share Special Edu. Needs	0.138	0.141	0.137	0.004	0.004
	[0.345]	[0.349]	[0.344]	(0.022)	(0.018)
School Size	47.681	51.619	47.059	4.560	10.951
	[25.941]	[22.032]	[26.453]	(6.325)	(4.371)
Panel B: Cohort 2					
Age-7 Test	15.935	15.612	15.988	-0.376	0.039
	[3.444]	[3.384]	[3.451]	(0.292)	(0.292)
Share Free School Meals	0.233	0.247	0.231	0.016	0.023
	[0.423]	[0.432]	[0.422]	(0.028)	(0.028)
Gender: Male	0.507	0.480	0.511	-0.031	-0.057
	[0.500]	[0.500]	[0.500]	(0.020)	(0.024)
Share Special Edu. Needs	0.134	0.132	0.134	-0.002	0.024
	[0.340]	[0.338]	[0.341]	(0.023)	(0.023)
School Size	50.236	50.759	50.151	0.608	9.332
	[31.747]	[21.904]	[33.073]	(7.375)	(4.782)
Pair FX					X

Notes: Obs. in Panel A/B: 6,436/6,298. Standard deviations of variables shown in square parenthesis in columns 1-3. Standard errors clustered at the school level shown in round parenthesis in columns 4-5.

Table A.2: Analysis of Individual-Level Dropout in Treated Schools

	(1) Treated School	(2) Treated Students	(3) Untreated Students	(4) (2)-(3)	(5) (2)-(3)
Panel A: Cohort 1					
Age-7 Test	15.614	15.584	15.668	-0.083	-0.351
	[3.539]	[3.544]	[3.532]	(0.282)	(0.247)
Share Free School Meals	0.237	0.221	0.268	-0.047	-0.049
	[0.425]	[0.415]	[0.443]	(0.027)	(0.029)
Gender: Male	0.502	0.505	0.496	0.009	0.085
	[0.500]	[0.500]	[0.500]	(0.017)	(0.035)
Share Special Edu. Needs	0.139	0.130	0.157	-0.027	-0.034
	[0.346]	[0.336]	[0.364]	(0.019)	(0.026)
School Size	46.896	43.329	53.508	-10.179	0.002
	[24.641]	[23.041]	[26.112]	(4.793)	(0.002)
Panel B: Cohort 2					
Age-7 Test	15.823	15.870	15.728	0.142	0.362
	[3.455]	[3.385]	[3.593]	(0.288)	(0.429)
Share Free School Meals	0.240	0.235	0.250	-0.015	-0.035
	[0.427]	[0.424]	[0.433]	(0.030)	(0.032)
Gender: Male	0.504	0.513	0.486	0.027	-0.021
	[0.500]	[0.500]	[0.500]	(0.019)	(0.025)
Share Special Edu. Needs	0.126	0.127	0.125	0.002	0.007
	[0.332]	[0.333]	[0.330]	(0.022)	(0.036)
School Size	47.422	43.863	54.618	-10.755	0.000
	[23.257]	[20.308]	[26.903]	(5.586)	(0.000)
Pair FX					X

Notes: Obs. in Panel A/B: 3,153/3,176 of all students in schools that participated in the study. Standard deviations of variables shown in square parenthesis in columns 1-3. Standard errors clustered at the school level shown in round parenthesis in columns 4-5.

Table A.3: IV-diff-in-diff analysis

	(1)	(2)	(3)	(4)	(5)
	ITT	School LATE		Student LATE	
Panel A: Cohort 1					
Test Score	0.444	0.564	0.022	0.677	0.027
	(1.059)	(1.340)	(0.053)	(1.612)	(0.064)
Maths	1.204	1.531	0.055	1.838	0.066
	(1.225)	(1.547)	(0.055)	(1.873)	(0.067)
Reading	-0.317	-0.404	-0.015	-0.484	-0.018
	(1.040)	(1.328)	(0.049)	(1.593)	(0.058)
SPAG	-0.174	-0.221	-0.008	-0.265	-0.010
	(1.193)	(1.519)	(0.057)	(1.821)	(0.068)
Science	-0.008	-0.046	-0.072	-0.055	-0.086
	(0.021)	(0.036)	(0.056)	(0.043)	(0.067)
First Stage		0.786		0.655	
		(0.042)		(0.038)	
Panel B: Cohort 2					
Test Score	0.774	0.976	0.039	1.153	0.046
	(1.176)	(1.473)	(0.058)	(1.751)	(0.069)
Maths	1.322	1.666	0.059	1.968	0.070
	(1.392)	(1.743)	(0.062)	(2.078)	(0.074)
Reading	0.227	0.286	0.010	0.338	0.012
	(1.121)	(1.410)	(0.051)	(1.668)	(0.061)
SPAG	-0.788	-0.993	-0.037	-1.173	-0.044
	(1.448)	(1.448)	(0.054)	(1.707)	(0.064)
First Stage		0.794		0.672	
		(0.042)		(0.038)	
Standardised			X		X

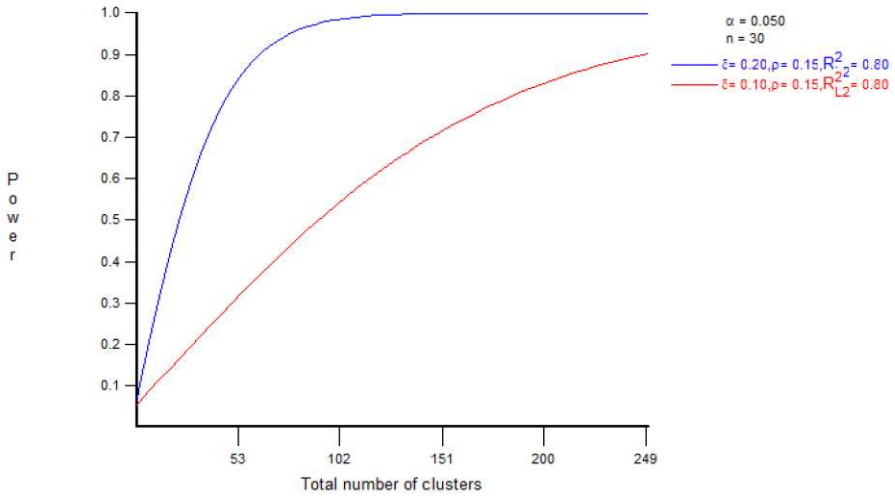
Notes: This tables shows difference-in-difference estiamtes for specification 4 in column (1) and combined DID-IV estimates at the school and student level. For both cohorts, included are all years from 2008/9 up to the start of the trial in 2012/13 as control years. Pair-FX, Age-7 test scores and student demographics are always included as controls. Standard errors clustered at the school level in parenthesis.

Table A.4: Heterogeneity, total effects by subgroup

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Test Scores		Maths		Reading		SPAG	
Panel A: Cohort 1								
Share Free School Meals	0.175	0.007	1.052	0.037	-0.702	-0.026	-1.212	-0.044
	(1.173)	(0.046)	(1.348)	(0.048)	(1.239)	(0.045)	(1.495)	(0.054)
ESL	-0.065	-0.003	1.659	0.059	-1.789	-0.065	3.063	0.111
	(1.517)	(0.060)	(1.768)	(0.063)	(1.708)	(0.062)	(2.061)	(0.075)
Minority	0.479	0.019	2.452	0.087	-1.495	-0.054	3.738	0.135
	(1.469)	(0.058)	(1.692)	(0.060)	(1.658)	(0.060)	(1.943)	(0.070)
Low age-7 test score	1.252	0.049	1.921	0.068	0.584	0.021	2.684	0.097
	(1.360)	(0.054)	(1.505)	(0.054)	(1.470)	(0.053)	(1.539)	(0.056)
Gender: male	0.844	0.033	1.498	0.053	0.190	0.007	0.917	0.033
	(1.048)	(0.041)	(1.225)	(0.044)	(1.057)	(0.038)	(1.247)	(0.045)
Panel B: Cohort 2								
Share Free School Meals	0.561	0.022	1.175	0.041	-0.053	-0.002	-0.383	-0.014
	(1.299)	(0.051)	(1.473)	(0.052)	(1.403)	(0.051)	(1.514)	(0.053)
ESL	1.788	0.070	3.207	0.113	0.369	0.013	1.585	0.056
	(1.806)	(0.071)	(2.174)	(0.077)	(1.818)	(0.066)	(2.223)	(0.079)
Minority	1.845	0.073	2.721	0.096	0.969	0.035	1.053	0.037
	(1.517)	(0.060)	(1.746)	(0.062)	(1.658)	(0.060)	(1.748)	(0.062)
Low age-7 test score	0.465	0.018	1.611	0.057	-0.681	-0.025	0.172	0.006
	(1.447)	(0.057)	(1.712)	(0.060)	(1.526)	(0.055)	(1.596)	(0.056)
Gender: male	-0.007	-0.000	0.038	0.001	-0.052	-0.002	-0.523	-0.018
	(1.281)	(0.050)	(1.472)	(0.052)	(1.267)	(0.046)	(1.320)	(0.047)
Standardised		X		X		X		X

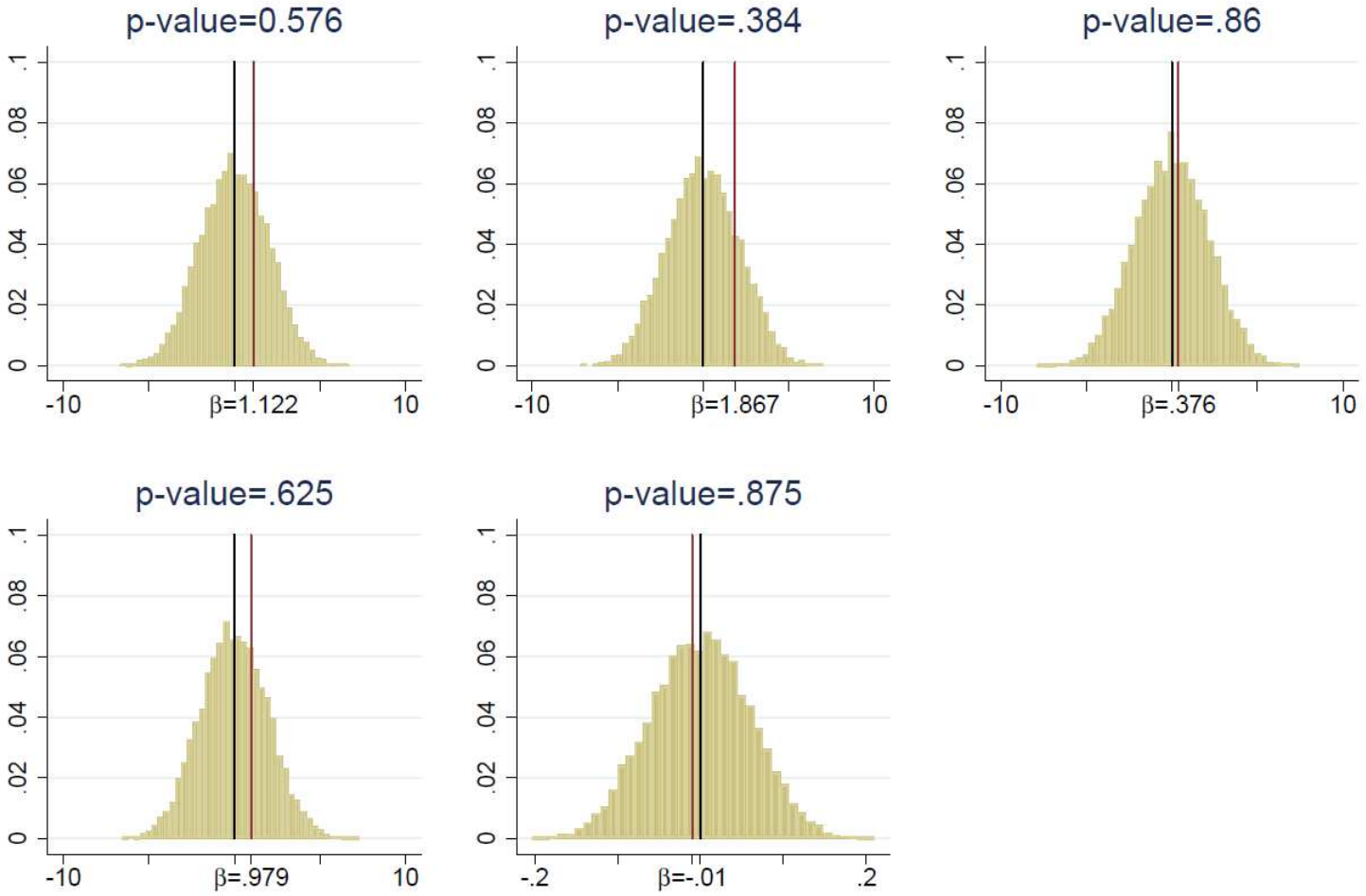
Notes: This tables shows estimates for total effects (main effect and interaction with respect to characteristic) for age-11 outcomes in test scores, maths, reading and SPAG scores in columns 1 to 8. Results presented in columns 3 to 8 were not part of the pre-registered analysis plan of the intervention. All specfications include pair FX and age-7 test scores as controls. Standard errors clustered at school level in parenthesis.

Figure A.1: Power calculations, pre-trial



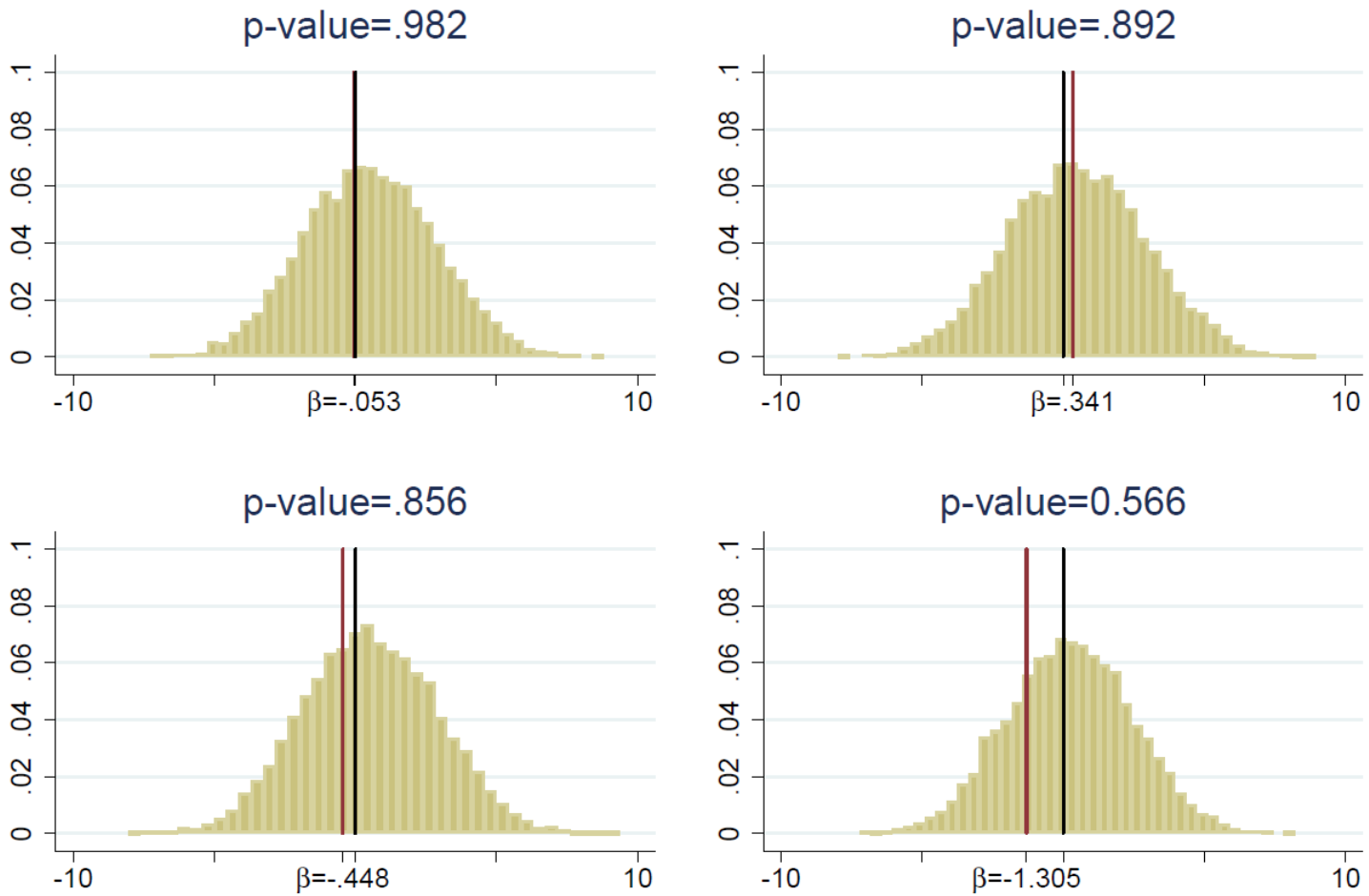
Notes: Blue line indicates power with effect size of 0.2 s.d., red line effect size of 0.1 s.d.

Figure A.2: Simulated fisher exact p-values, cohort 1



Notes: To obtain these distributions, treatment status was randomly assigned within school pairs. 10,000 simulations each. This is for Table 3, Panel A.

Figure A.3: Simulated fisher exact p-values, cohort 2



Notes: To obtain these distributions, treatment status was randomly assigned within school pairs. 10,000 simulations each. This is for Table 3, Panel B.